A Hybrid Approach of Pattern Extraction and Semi-supervised Learning for Vietnamese Named Entity Recognition

Duc-Thuan Vo and Cheol-Young Ock

Natural Language Processing Lab,
School of Computer Engineering and Information Technology,
University of Ulsan, Korea
thuanvd@gmail.com, okcy@ulsan.ac.kr

Abstract. Requiring a large hand-annotated corpus in supervised learning of contemporary Vietnamese Named Entity Recognition researches is challenging. We therefore propose a hybrid approach of pattern extraction and semi-supervised learning. Applied rule-based method helps generating patterns automatically. Part-of-speech tagger, lexical diversity and chunking are explored to define rules in pattern extractions which are used for identifying potential named entities. Semi-supervised learning trains a small amount of seed named entities to categorize named entities in extracted patterns. In experiments, our approach shows good increasing the system accuracy with others in Vietnamese.

Keywords: named entity recognition, part-of-speech, chunking, rule-based, pattern extraction, semi-supervised learning.

1 Introduction

Named Entity Recognition (NER) is now considered as a fundamental for many natural language processing tasks such as information retrieval, machine translation, information extraction and question answering. In information extraction NER is employed to classify text's atomic elements into predefined categories such as name of person, organization, location, expression of time, quantity, monetary value, and percentage. In fact, most NER system studies have been structured by taking an unannotated block of text. For example, a given sentence "Tim Cook has been CEO of Apple Inc. since 2011." will produce an annotated block of text such as:

<PER>Tim Cook</PER>has been CEO of <ORG> Apple Inc.</ORG> since <Date>2011</Date>.

Recently, NER system studies are expanded and concentrate on popular languages as English, France, Spanish, Japan and etc. But researching NER in Vietnamese is quite new. Tri et al. [16] performed NER based on Support Vector Machine (SVM). The system VN-KIM IE [8] built ontology and then applied into Japa grammars to define target named entities in web. And Nguyen et al. [7] employed rule-based approach for

N.-T. Nguyen et al. (Eds.): ICCCI 2012, Part I, LNAI 7653, pp. 83–93, 2012.

Jape grammars plug-in Gate framework for NER. However, these approaches require a large hand-annotated corpus that takes large amount of example data and expertise to label or annotate training data. The problem can be solved by semi-supervised learning methods that trains data in only initial set of small labeled data and makes recognition unlabelled data, then iteratively attempts to create an improved model using predictions of previously generated model as plus the original labeled data.

We thus propose a hybrid approach of pattern extraction and semi-supervised learning for NER in Vietnamese text. The first part is defining rules to automatically generate patterns, called predefined target entities. Next, semi-supervised learning method is employed to train data with a small initial labeled dataset, and then categorize entities in extracted patterns for NER. For preprocessing, we employ corpuses in Vietnamese Treebank of VLSP project[1] due to exclude errors on POS tagging and chunking parsing as input for experimental setting.

The rest of the paper is structured as follows. Section 2 presents related work that refers to named entities recognition system. In Section 3 Vietnamese's features are introduced as a base for defining rules. And in Section 4 rule-based method is employed for pattern extraction, and Self-training algorithm is applied to Semi-supervised learning. The next Section, experiments results applied in Vietnamese corpus are showed and discussed. The last Section ends with conclusion and future work.

2 Related Work

Supervised learning in NER takes a large amount of example data and expertise to label or annotate on the training data. The method defines important entities and classifies them into predefined categories as person, organization, location, time and so on for training. NER systems have been developed through using supervised learning methods as Decision Tree [3], Maximum Entropy Model [17], and Support Vector Machine [14] which achieve high result. Annotating such corpuses of these researches requires great human effort. However, unsupervised learning method can learn without examples. The method refers to the problem in trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution.

NER researches in Vietnamese, Tri et al. [16] performed SVM based on NER and compared with Conditional Random Fields. VN-KIM IE [8], the system has recognized name entities in Vietnamese web pages by indentifying their classes and addresses in knowledge. They used ontology as knowledge base containing 370 classes and 115 properties with over 120,000 entities. Nguyen et al. [7] presented a rule-based system, where the rules are incrementally created while annotating a named entity corpus. The rule-based system will bootstrap the annotation process and the corpus is already pre-annotated. However, the limitation of these systems are in changing named entities domain or solving of ambiguity two different names refer to

[1] http://vlsp.vietlp.org:8080